

# Chapitre 1 : Motivations et régression linéaire simple

## 1 Motivations

Les premiers outils que vous avez vu en **statistique inférentielle** permettent d'obtenir de l'information sur la valeur de paramètres liés à des **populations** à partir d'un **échantillon**. Nous allons voir dans ce cours comment :

- prédire la valeur d'une variable  $Y$  au niveau d'individus, dite « **variable à expliquer** », « output » ou « sortie », (ou maladroitement « variable dépendante »)
- en s'appuyant sur les valeurs d'autres variables  $X_1, \dots, X_p$  observées sur le même individu, appelées « **covariables** », « variables explicatives », « régresseurs » (ou maladroitement « variables indépendantes »)

Pour cela, nous allons devoir comprendre la régularité du phénomène qui lie la variable explicative aux covariable au niveau de la population à partir d'un

échantillon d'individus, puis descendre au niveau du nouvel individu pour faire la prédiction. Par nature, c'est un problème compliqué puisqu'il y a deux niveaux, l'individu et la population, et plusieurs variables sur chacun des individus.

Le **modèle linéaire** est la façon la plus simple de répondre à cet objectif compliqué. Nous verrons qu'il repose sur des **hypothèses très fortes** sur le lien entre  $Y$  et les covariables au niveau de la population. Mais ces hypothèses permettent de construire une théorie complète d'inférence et de prédiction qui est la pierre de base du « **machine learning** » ou apprentissage automatique en français.

Il y a de très nombreuses choses à comprendre, et **le contenu de ce cours sera dense**. Voici le plan général :

**CM 1** Aujourd'hui : motivations et régression linéaire simple

**TD/TP 1** Premières manipulations mathématiques et informatiques

**CM 2** Régression linéaire multiple

**TD/TP 2** Mise en œuvre et interprétation, comprendre le théorème de Cochran

**CM 3** Hypothèses et diagnostic, interactions, plan d'expérience, pénalisation  $L^2$  (ridge)

**CM 4 QCM d'évaluation**, imputation, choix de modèle, pénalisation  $L^1$  (Lasso)

**TD/TP 3** Mise en œuvre du choix de modèle et du Lasso, calculs théoriques  
autour des  $C_p$ , AIC et BIC

**CM 5** Introduction aux GLM avec la régression logistique et la régression de Poisson

**TD/TP 4** Mise en œuvre de ces modèles, calcul formel de la vraisemblance et de son gradient

**Epreuve écrite** 1h30 sur feuille

Certains exercices de TP seront évalués par compte-rendu. Ils seront réalisés avec R ou SAS.

Dans les outils du statisticien, vous aurez besoin :

- d'algèbre linéaire : calcul matriciel, projection dans des espaces euclidiens
- d'analyse : calcul de gradient de fonctions de plusieurs variables
- du calcul des probabilités (variables aléatoires, vecteurs aléatoires, espérance, variance, corrélation, matrice de variance-covariance,...)
- de la notion d'indépendance, de conditionnement et déconditionnement
- de la loi normale ou gaussienne univariée, et multivariée

- des notions d'estimateurs et d'erreur d'inférence, d'intervalles de confiance et un peu de test

## 1.1 Prédire en l'absence de covariable

Comment peut-on prédire la valeur d'une variable  $Y$  en l'absence de toute autre information sur l'individu ?

**Etape 1** On collecte un échantillon  $y_1, \dots, y_n$  issu de la population

**Etape 2** On tire de l'information sur la population à partir de cet échantillon

**Etape 3** On utilise cette information pour prédire la valeur de  $Y$  sur un nouvel individu, en faisant éventuellement attention à l'erreur possible à l'étape 2

On veut prédire la valeur de  $Y$  sur un nouvel individu à l'aide d'**une seule valeur**, caractéristique de la population, notée  $\theta$ . Par exemple, on peut prendre la moyenne de la population à cause du résultat ci-dessous.

**Théorème 1.** Soit  $Y$  une variable aléatoire réelle  $L^2$ . Soit  $d$  la fonction définie

↑ de carré intégrable (i)  $\mathbb{E}(Y^2) < +\infty$

pour tout  $\theta \in \mathbb{R}$  par

$$d(\theta) = \mathbb{E}\left((Y - \theta)^2\right).$$

Cette fonction admet un minimum global en  $\theta = \mathbb{E}(Y)$ . Et  $d(\mathbb{E}(Y)) = \text{Var}(Y)$ .

Soit  $y_1, \dots, y_n$  une série de  $n$  nombres. Soit  $\phi$  la fonction définie pour tout  $\theta \in \mathbb{R}$  par

$$\phi(\theta) = \sum_{i=1}^n (y_i - \theta)^2.$$

Cette fonction admet une unique minimum global en  $\theta = \bar{y}$ .

Interprétation : le meilleur résumé numérique  $\theta$  d'une v.a.  $Y$  au sens de **l'erreur des moindres carrés** est son espérance  $\mathbb{E}(Y)$ .

Ainsi, à l'étape 2, on peut chercher à estimer  $\theta = \mathbb{E}(Y)$ , la moyenne de la population par

$$\hat{\theta} = \frac{Y_1 + \dots + Y_n}{n},$$

où  $Y_1, \dots, Y_n$  sont des v.a. qui modélisent l'échantillon, que l'on suppose indépendant, tiré suivant la même loi (=dans la même population).

Pour réaliser le programme de l'étape 3, on doit alors répondre par la valeur observée de  $\hat{\theta}$  quel que soit le nouvel individu. Pour comprendre l'erreur commise, on introduit une nouvelle variable aléatoire  $Y$ , indépendante de l'échantillon, tirée dans la même loi que les  $Y_i$ .

**Proposition 2.** Soit  $Y$  une v.a.  $L^2$  d'espérance  $\theta$  et de variance  $\sigma^2$ . On a

$$\mathbb{E}\left((Y - \theta)^2\right) = \text{Var}(Y) = \sigma^2.$$

Et

$$\mathbb{E}\left((Y - \hat{\theta})^2\right) = \text{Var}(Y) + \frac{\text{Var}(Y)}{n} = \sigma^2 \left(1 + \frac{1}{n}\right)$$

Le premier terme vient de la **variabilité entre individus** de la population. Le second terme vient de **l'erreur d'inférence sur  $\theta$** . Ce terme correctif  $+1/n$  est dû à la **propagation de l'incertitude** sur la valeur de  $\theta$  dans la prédiction de  $Y$ .

Enfin, on peut fournir une réponse plus détaillée à l'étape 3, en renvoyant non pas une seule valeur,  $\hat{\theta}$ , mais un intervalle de valeurs, avec une certaine probabilité ou couverture. Pour cela, il faut faire **une hypothèse plus forte sur la distribution de  $Y$**  au sein de la population : on suppose que  $Y \sim \mathcal{N}(\theta, \sigma^2)$ .

**Proposition 3.** *Sous ces hypothèses de lois gaussiennes, on peut estimer  $\sigma^2$  par l'estimateur usuel  $\hat{\sigma}^2 = (n-1)^{-1} \sum_i (Y_i - \hat{\theta})^2$  et obtenir l'intervalle de confiance :*

$$\forall \alpha \in ]0; 1[, \quad \mathbb{P} \left( \theta \in \left[ \hat{\theta} \pm \Phi_{n-1}^{-1}(1 - \alpha/2) \frac{\hat{\sigma}}{\sqrt{n}} \right] \right) = 1 - \alpha$$

où  $\Phi_{n-1}^{-1}$  est la fonction quantile de la loi de Student  $t_{n-1}$ .

NB : si  $n$  est grand,  $t_{n-1} \approx \mathcal{N}(0; 1)$ .

Pour un nouvel individu  $Y$ , indépendant des observations, on a un résultat équivalent.

**Proposition 4.** *Sous les mêmes hypothèses, et avec les mêmes notations, on a*

$$\forall \alpha \in ]0; 1[, \quad \mathbb{P} \left( Y \in \left[ \hat{\theta} \pm \Phi_{n-1}^{-1}(1 - \alpha/2) \hat{\sigma} \sqrt{1 + \frac{1}{n}} \right] \right) = 1 - \alpha.$$

Dans les deux cas, on notera :

- que la loi de Student apparaît car on remplace une variance au niveau de la population  $\sigma^2$  par son estimateur sans biais

— que les résultats de la Prop 2 se trouve directement transcrits en facteur du quantile de cette loi.

Le second intervalle, qui tient compte à la fois de la variabilité individuelle et de l'erreur d'inférence est un intervalle de prédiction. Quand  $n \rightarrow \infty$ , la largeur de l'intervalle de confiance sur  $\theta$  tend vers 0 (l'information est maximale). Mais l'erreur individuelle sur une prédiction pour un nouvel individu subsiste.

Au final, on peut imaginer d'autres modèles, qui supprime l'hypothèse gaussienne, mais cela dépasse les objectifs de ce cours.

Le dernier modèle peut s'écrire de différentes façons pour un individu  $Y$  de la population.

1.  $Y \sim \mathcal{N}(\theta, \sigma^2)$

2.  $Y = \theta + \varepsilon$ , où  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

Cette dernière variable,  $\varepsilon$ , jamais observée car  $\theta$  est inconnu, s'appelle **l'erreur**.

## 2 Régression linéaire simple

On s'intéresse à une population d'individus sur lesquels on observe maintenant deux variables,  $X$  et  $Y$ . On fait l'hypothèse que, pour un individu,

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad (1)$$


où  $\beta_0, \beta_1$  sont deux paramètres définis au niveau de la population qui caractérisent le lien entre  $X$  et  $Y$ , et  $\varepsilon$  est indépendant de  $X$ , d'espérance nulle. C'est une équation du même style que la forme 2 du modèle sans covariable.

**Proposition 5.** *Sous ces hypothèses, on a*

$$\beta_0 = \mathbb{E}(Y) - \beta_1 \mathbb{E}(X) \quad \text{et} \quad \beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}.$$

Cette proposition montre que  $\beta_0$  et  $\beta_1$  sont bien des paramètres définis au niveau de la population.

L'indépendance entre la covariable  $X$  et l'erreur  $\varepsilon$  permet de dire que l'on a décomposé la variable  $Y$  en somme de deux termes :

- un terme  $\beta_0 + \beta_1 X$  qui ne dépend d'un individu à l'autre que de la valeur de  $X$ ,
- un terme  $\varepsilon$  qui est indépendant de  $X$ , mais varie aussi d'un individu à l'autre.

Vue la décomposition de  $Y$  en deux termes indépendants, on a

**Proposition 6.**

$$\text{Var}(Y) = \beta_1^2 \text{Var}(X) + \sigma^2.$$

On peut donc comparer la variabilité de l'erreur avec la variabilité totale par la fraction de variance de  $Y$  expliquée par  $X$  avec :

$$R_{\text{pop}}^2 = \frac{\beta_1^2 \text{Var}(X)}{\text{Var}(Y)} = \text{Cor}^2(X, Y)$$

où l'on a utilisé la proposition ?? pour la dernière égalité. Ce nombre, entre 0 et 1, vaut 1 si et seulement si  $\varepsilon = 0$ . Et il faut 0 si et seulement si  $Y$  est indépendante de  $X$ .

Dans le même esprit que le théorème 1 en l'absence de covariable, on a

**Théorème 7.** *Sous les hypothèses qui précèdent, soit  $d$  la fonction définie pour*

tout  $b_0, b_1$  par

$$d(b_0, b_1) = \mathbb{E}\left((Y - (b_0 + b_1 X))^2\right).$$

Elle admet un unique minimum global en  $(b_0, b_1) = (\beta_0, \beta_1)$ .

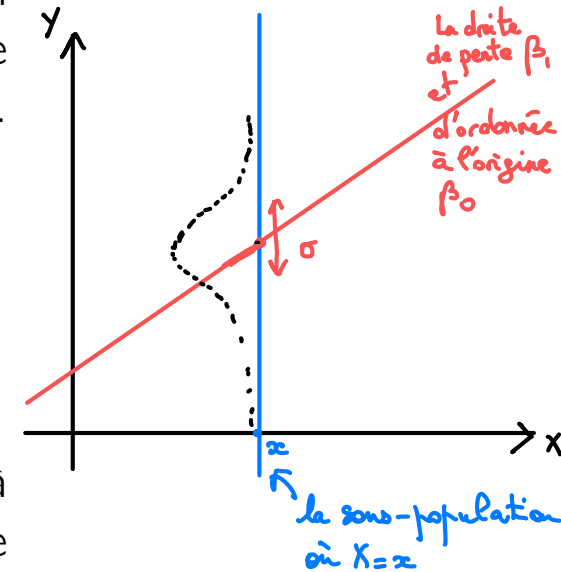
Dans la présentation ci-dessus, il manque un terme important dans la modélisation qui définit la façon dont  $\varepsilon$  varie d'un individu à l'autre :  $\sigma^2 = \text{Var}(\varepsilon)$ . Ainsi, on a formulé un modèle à trois paramètres,  $\theta = (\beta_0, \beta_1, \sigma^2)$ .

On peut ajouter une hypothèse sur la distribution de l'erreur :  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ . Il s'agit alors du modèle de régression linéaire simple gaussien. La forme 1 de ce modèle est plus compliquée à écrire que dans le cas où il n'y a pas de covariable. Ici, elle s'écrit sous la forme de la loi conditionnelle ci-dessous.

**Proposition 8.** Sous les hypothèses du modèle linéaire simple gaussien, on a

$$\left[Y \mid X = x\right] \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2).$$

Ce qui s'interprète comme : dans la sous-population où la variable  $X$  est fixée à la valeur  $x$ , la variable  $Y$  suit une loi normale centrée en  $\beta_0 + \beta_1 x$  et de variance  $\sigma^2$ .



### 3 Inférence

Pour estimer  $\theta$ , on a recours à un échantillon indépendant de  $n$  individus tirés dans la population, sur lequel on observe les valeurs des deux variables simultanément. Ainsi, notre jeu de données est composé des  $n$  paires  $(x_i, y_i)$ ,  $i = 1, \dots, n$ .

Pour modéliser ces données, on introduit donc  $n$  paires indépendantes de v.a.  $(X_i, Y_i)$  qui vérifient :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{Var}(\varepsilon_i) = \sigma^2, \quad X_i \perp\!\!\!\perp \varepsilon_i.$$

Il faut noter ici qu'il existe **deux types de situations** :

- des cas où on peut **choisir la valeur de  $X$**  au moment où on collecte les données,
- et des cas où **les valeurs de  $X$  sont subies** au moment où on tire un individu, et où on a donc un échantillon où les valeurs de  $X$  sont représentative de leur distribution dans la population.

Tous les résultats théoriques dans la suite sont établis conditionnellement aux valeurs observées de  $X$  dans l'échantillon, donc ces deux situations sont équivalentes.

En s'inspirant du théorème 1, on peut introduire l'estimateur des moindres carrés pour inférer  $\beta_0$  et  $\beta_1$ . Autrement dit, on cherche  $(\hat{\beta}_0, \hat{\beta}_1)$  qui minimisent la fonction

$$d(b_0, b_1) = \sum_{i=1}^n \left( Y_i - (b_0 + b_1 X_i) \right)^2.$$

On écrit

$$(\hat{\beta}_0, \hat{\beta}_1) = \operatorname{argmin}_{b_0, b_1} d(b_0, b_1).$$

**Proposition 9.** *Sous les hypothèses qui précèdent, on a*

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} \quad \text{et} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Les formules d'inférences sont explicites. De plus, il est d'usage d'estimer  $\sigma^2$  avec

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \left( Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)^2$$

qui a de bonnes propriétés.

Voici, avec ces hypothèses, les résultats que l'on peut démontrer.

**Théorème 10.** *Sous les hypothèses qui précèdent, les estimateurs  $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$  sont sans biais conditionnellement à  $(X_1, \dots, X_n)$  (noté  $X_{1:n}$ ). C'est-à-dire*

$$\mathbb{E}(\hat{\theta} | X_{1:n} = x_{1:n}) = \theta$$

*Et les  $\hat{\beta}_i$  sont et de variance minimale conditionnellement à  $(X_{1:n})$ . Et, toujours conditionnellement à  $X_{1:n}$ ,  $\hat{\sigma}^2$  est indépendant de  $(\hat{\beta}_0, \hat{\beta}_1)$ .*

La seconde partie est difficile à démontrer et doit être admise.

En revanche, on peut montrer le résultat suivant.

**Proposition 11.** *Sous les hypothèses précédente, on a*

$$\text{Var}(\hat{\beta}_1 | X_{1:n} = x_{1:n}) = \sigma^2 \frac{1}{\sum_i (x_i - \bar{x})^2}$$

et

$$\text{Var}(\hat{\beta}_0 | X_{1:n} = x_{1:n}) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right).$$

*De plus, si  $x \in \mathbb{R}$ ,*

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x | X_{1:n} = x_{1:n}) = \sigma^2 \left( \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right)$$

Ce dernier terme,  $\hat{\beta}_0 + \hat{\beta}_1 x$  est l'estimateur de  $\beta_0 + \beta_1 x$ , l'espérance de  $Y$  sur la sous-population où  $X = x$ . C'est le terme que l'on utilise si on veut **prédire la valeur de**  $Y$  pour un nouvel individu, indépendant de l'échantillon, pour lequel on a observé que  $X = x$ .

Une fois  $X = x$  connu, cette espérance conditionnelle  $\beta_0 + \beta_1 x$  est un paramètre lié à la sous-population où  $X = x$ . La dernière variance est donc l'erreur d'inférence sur ce paramètre. Comme dans le cas où il n'y a pas de covariable, on a maintenant

**Proposition 12.** *Si on considère un nouvel individu  $(X, Y)$ ,*

$$\text{E} \left( \text{Var} \left( Y - \hat{\beta}_0 - \hat{\beta}_1 X \right) \middle| X = x \right) = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right).$$

Dans cette erreur de prédiction de  $Y$  (inconnue) par  $\hat{\beta}_0 + \hat{\beta}_1 x$  (connu), on constate à nouveau que l'on tient compte :

- de la variabilité de  $\varepsilon$  d'un individu à l'autre avec  $\sigma^2(1 +$
- de l'erreur d'inférence sur les vraies valeurs  $\beta_0$  et  $\beta_1$ , erreur que l'on propage ici au niveau de la prédiction.

**Le cas gaussien.** On peut maintenant ajouter l'hypothèse que  $\varepsilon \sim \mathcal{N}(0; \sigma^2)$ , c'est-à-dire

$$\left[ Y \middle| X = x \right] \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2).$$

$$X_{1:n} = (x_1, \dots, x_n)$$

Alors, la loi de l'échantillon, conditionnellement aux valeurs observées des covariables est une loi gaussienne multivariée :

$$\left[ Y_{1:n} \middle| X_{1:n} = x_{1:n} \right] \sim \mathcal{N}_n(\beta_0 \mathbf{1} + \beta_1 x_{1:n}, \sigma^2 I_n),$$

où  $\mathbf{1}$  est le vecteur dont les coordonnées sont toutes égales à 1 et  $I_n$  est la matrice identité d'ordre  $n$ .

Dans cette situation, la fonction de vraisemblance conditionnelle est :

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n \left\{ \varphi \left( \frac{y_i - \beta_0 - \beta_1 x_i}{\sigma} \right) \right\}$$

où  $\varphi$  est la densité de la loi normale centrée réduite. Un rapide calcul montre que, la log-vraisemblance vaut

$$\ell(\beta_0, \beta_1, \sigma^2) = \log L(\beta_0, \beta_1, \sigma^2) = -\frac{1}{2\sigma^2} \sum_i \left( y_i - \beta_0 - \beta_1 x_i \right)^2 - \frac{n}{2} \log(\sigma^2).$$

**Proposition 13.** *L'estimateur du maximum de vraisemblance est donné par*

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} \quad \text{et} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

*Autrement, ce sont les estimateurs des moindres carrées pour  $\beta_0$  et  $\beta_1$ . De plus, l'estimateur du maximum de vraisemblance pour le dernier paramètre est*

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \left( Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)^2.$$

*Ce dernier estimateur est différent de celui introduit précédemment, et est biaisé. L'estimateur  $\hat{\sigma}^2$  est non biaisé (conditionnellement aux valeurs de  $X$ ) et on le préfère.*

Enfin, comme dans le cas où il n'y a pas de covariable, on peut quantifier les incertitudes par des intervalles :

- un intervalle de confiance sur  $\beta_0 + \beta_1 x$ , qui est un paramètre défini au niveau de la sous-population où  $X = x$ ,
- un intervalle de prédiction sur  $Y$  restreint à la sous-population où  $X = x$ .

**Proposition 14.** *On note  $\varphi_{n-2}^{-1}$  la fonction quantile de la loi de Student  $t_{n-2}$ .*

Sous les hypothèses gaussiennes, on a, pour tout  $\alpha \in ]0;1[$ ,

$$\mathbb{P} \left( \beta_0 + \beta_1 X \in \left[ \hat{\beta}_0 + \hat{\beta}_1 X \pm \varphi_{n-2}^{-1} \left(1 - \frac{\alpha}{2}\right) \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} \right] \middle| X_{1:n} = x_{1:n}, X = x \right) = 1 - \alpha.$$

Et, pour tout  $\alpha \in ]0;1[$ ,

$$\mathbb{P} \left( Y \in \left[ \hat{\beta}_0 + \hat{\beta}_1 X \pm \varphi_{n-2}^{-1} \left(1 - \frac{\alpha}{2}\right) \hat{\sigma} \sqrt{\mathbf{1} + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} \right] \middle| X_{1:n} = x_{1:n}, X = x \right) = 1 - \alpha.$$

À nouveau ici, on voit que l'on a propagé l'incertitude (liée à la substitution des  $\beta_i$  par leur estimateur) à la prédiction de  $Y$  sachant  $X = x$ .

## 4 Quelques remarques supplémentaires

**Erreur n°1** Ne pas confondre le terme d'erreur  $\varepsilon_i = Y_i - \beta_0 - \beta_1 X$  qui n'est jamais observé, avec le terme

$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

que l'on peut calculer (puisque l'on a substitué les paramètres inconnus par leurs estimateurs). Ces derniers termes s'appellent les **résidus**.

Ces résidus sont utiles, puisqu'ils reconstruisent les valeurs non observées de  $\varepsilon_i$ . Mais les résidus ont tendance à être plus petits que les  $\varepsilon_i$  car les valeurs substituées  $\hat{\beta}_0$  et  $\hat{\beta}_1$  dépendent aussi des valeurs  $(X_i, Y_i)$  de l'échantillon.

[Pour s'en convaincre, il suffit de voir que  $\hat{\sigma}^2 = \sum e_i^2 / (n - 2)$  alors que l'on estimerait cette variance avec  $\sum \varepsilon_i^2 / (n - 1)$  si ces dernières variables étaient observables.]

**Erreur n°2** Le  $R^2_{\text{pop}}$  est un nombre intéressant au niveau de la population, puisqu'il mesure à quel point  $X$  permet de prédire  $Y$ . Il donne la part de variance de  $Y$  expliquée par la meilleure fonction affine de  $X$ .

Son estimateur est :

$$R^2 = \frac{\text{SSR}}{\text{SST}}, \text{ où}$$

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$SSE = \sum_{i=1}^n e_i^2, \quad \text{et}$$

$$SSR = \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y})^2$$

Ici,  $SST = \text{Total Sum of Square}$ ,  $SSR = \text{Sum of Square due to Regression}$  and  $SSE = \text{Sum of Squares Error}$  (nom maladroit car il s'agit de la somme des carrés des résidus !)

On a toujours  $SST = SSR + SSE$  et ces trois nombres sont positifs. Sur l'estimateur, on a à nouveau que le  $R^2$  est une évaluation du lien entre  $Y$  et sa prédiction linéaire.

**La qualité de l'ajustement du modèle aux données ne s'évalue pas avec  $R^2$ . C'est de la qualité des prédictions de  $Y$  avec  $X$  dont parlent ce  $R^2$  !**

**Erreur n°3** Dans le cas gaussien, il existe des tests statistiques pour les jeux d'hypothèses ci-dessous.

Jeu n°1 :

$$\mathcal{H}_0 : \beta_0 = 0 \quad \text{vs} \quad \mathcal{H}_1 : \beta_0 \neq 0.$$

Jeu n°2 :

$$\mathcal{H}_0 : \beta_1 = 0 \quad \text{vs} \quad \mathcal{H}_1 : \beta_1 \neq 0.$$

Leur mise en œuvre revient à vérifier que 0 est dans l'intervalle de confiance. Mais, dans 99,9% des cas, **ils sont sans intérêt** dans le sens où ils répondent, de façon péremptoire, à une question qui est mal posée.

La question de savoir si la covariable  $X$  a un intérêt pour prédire  $Y$  ne se traite pas avec le test du jeu d'hypothèses n°2, mais avec une question de choix de modèles, entre le modèle de la partie 1 de ce chapitre, et le modèle de régression linéaire simple.

**Erreur n°4** La régression linéaire gaussienne ne repose pas sur le fait que  $X$  et  $Y$  suivent des lois gaussiennes. L'hypothèse de loi gaussienne est mise sur la variable non observée  $\varepsilon$ , ou, de manière équivalente, sur la loi de  $Y$  sachant  $X = x$ , c'est-à-dire sur la distribution de  $Y$  dans la sous-population où  $X = x$ .

On peut s'assurer que cette hypothèse est réaliste en représentant la distribution empirique des résidus, avec, par exemple, un **histogramme des  $e_i$** . Si cet

histogramme a une forme de cloche, c'est presque gagné.

**Erreur n°5** En première approche, on peut être tenté de ne pas s'intéresser au paramètre  $\sigma^2$ , qui représente la variance de  $Y$  au sein de la sous-population où  $X = x$ . Cette variance ne dépend pas de  $x$ . Sans le dire explicitement, on a fait une hypothèse forte ici. Dans la régression linéaire,  $\text{Var}(Y|X = x)$  ne dépend pas de la valeur de  $x$  (de la sous-population que l'on regarde). On parle de modèles **homoscédastiques**. On peut s'affranchir de cette hypothèse et construire des modèles, mais qui sont alors beaucoup plus compliqués et dépassent de très loin le cadre de ce cours.

**Erreur n°6** La dernière hypothèse, dont nous n'avons pas discuté, est de l'indépendance de  $\varepsilon$  et  $\beta_0 + \beta_1 X$ . Pour vérifier qu'elle soit vraie, et vérifier l'hypothèse d'homoscédasticité, il est courant de représenter le nuage des points de coordonnées  $(\hat{\beta}_0 + \hat{\beta}_1 x_i, e_i)$ . On met donc sur l'axe des abscisses les valeurs prédites de  $Y$  pour les individus de l'échantillon, et sur l'axe des ordonnées les résidus. Ce points doivent être répartis aléatoirement, sans structure, dans un rectangle parallèle à l'axe des abscisses.

Une erreur commune est de mettre sur l'axe des abscisses les valeurs  $y_i$  au lieu de mettre les prédictions. Le graphique où les  $y_i$  sont sur l'axe des abscisses n'a aucun intérêt (ou presque).